

# Uncertainty in Online Experiments with Dependent Data: An Evaluation of Bootstrap Methods

Eytan Bakshy  
Facebook  
eytan@fb.com

Dean Eckles  
Facebook  
deaneckles@fb.com

## ABSTRACT

Many online experiments exhibit dependence between subjects and items. For example, in online advertising, observations that have a user *or* an ad in common are likely to be associated. Because of this, even in experiments involving millions of subjects, the difference in means between control and treatment outcomes can have substantial variance. Previous mathematical and simulation results demonstrate that not accounting for this dependence structure can result in confidence intervals that are too narrow and inaccurate hypothesis tests.

We examine how bootstrap methods that account for differing levels of dependence structure perform in practice. We use multiple real datasets describing user behaviors on Facebook — responses to ads, search results, and News Feed stories — to generate data for experiments in which there is no effect of the treatment on average and then estimate empirical Type I error rates for each method. Results are supplemented with realistic simulations based on the data. Accounting for dependence within a single type of unit (i.e. within-user dependence) is often sufficient to get reasonable error rates. But when experiments have effects, as one might expect in the field, accounting for multiple units with a multiway bootstrap can be necessary to get close to the advertised Type I error rates. This work provides guidance to experimenters on calibrating large-scale evaluation systems, and highlights the importance of analysis of inferential methods under conditions in which experiments have effects.

## Categories and Subject Descriptors

Mathematics of computing [Bootstrapping]:

## General Terms

Experimentation, Theory

## Keywords

bootstrapping, field experiments, A/B testing, A/A tests,

statistical inference, online advertising, random effects, dependent data

## 1. INTRODUCTION

Experiments conducted on the Internet, including those taking place in social media feeds, online advertising, and search, frequently involve millions to tens of billions of observations. This could lead to the perception that there is little uncertainty about experimental outcomes. However, treatment effects are often very small in absolute terms, so many observations can be required to distinguish them from noise. Furthermore, many Internet-scale datasets have a structure such that individual observations are not independent; rather, there is substantial dependence between observations of the same units. For example, consider an online advertising experiment in which there 1 million ad impressions, but these only include 1,000 distinct ads and 10,000 distinct users. Clearly, the effective sample size may be something substantially less than 1 million, and there can be substantial uncertainty about the difference in click-through rate (CTR) between the treatment and control.

Accounting for this dependence is important for statistical inference, including hypothesis testing and confidence interval estimation. Inferential procedures that neglect this dependence structure are expected to be anti-conservative: they will have higher Type I error rates than expected and, e.g., “95%” confidence intervals will include the true value less than 95% of the time.

High false positive rates have substantial managerial consequences. For example, experiments using one popular experimentation platform at Facebook compare, on average, 3.7 non-control conditions. With four comparisons and nominal Type I error rate  $\alpha = 0.05$ , there should be a  $1 - (1 - 0.05)^4 = 18.5\%$  chance that at least one condition would be significant under the null hypothesis (i.e., one in 5.4 experiments with no effects may yield at least one significant condition). But if the true Type I error was considerably higher, say  $\alpha = 0.2$ , one would have a  $1 - (1 - 0.2)^4 = 59\%$  chance of having falsely rejected the null hypothesis. Given that many experiments involve comparing multiple outcomes (i.e., metrics), in practice the resulting effects on decision making can be worse than this suggests: not only can there be errors in comparisons of the primary outcome, but incorrectly rejecting the null for some other outcome might delay or prevent the launch of a change that is otherwise beneficial.

This paper describes sources and consequences of dependence in common applications of experimentation to Inter-

net services. We posit a general data generating process and illustrate how experimental assignment procedures and common effects of units (e.g., users and ads) affect the true uncertainty about experimental comparisons. We then evaluate independent, one-way, and multiway bootstrap methods for computing confidence intervals using null experiments (“A/A tests”) on three real datasets from Facebook: clicks on advertisements, search results, and content in the News Feed. We additionally modify these datasets to simulate systematic imbalance in items across conditions, as would result from changes to CTR prediction or ranking models. To examine performance under additional deviations from treatment effects, we conduct simulations using a realistic probit random effects model.

Our primary contribution is providing guidance about when accounting for dependence among observations is most important: while previous work has shown that neglecting all dependence structure results in massive overconfidence, less work has examined how accounting for some sources of dependence, but not others, affects inference in practice. We conclude that analysts should certainly use an inferential procedure that accounts for dependence among observations of the units assigned to conditions (e.g., users), but that whether not additionally accounting for secondary units (e.g., ads, search results, links) makes for misleading inference is more likely to depend on the specific (usually partially unknown) deviations from the sharp null hypothesis that the experiment has no effects.

The literature on routine Internet-scale experimentation (e.g. [8, 14]) largely does not address such questions about statistical inference. Some authors [14] suggest conducting null experiments to evaluate one’s experimentation tools, but little is said about how these should be conducted and exactly what problems they can detect. We intend that, in addition to our results, this paper provides a blueprint for other experimenters who wish to evaluate and choose among inferential procedures in their own settings.

## 2. DEPENDENCE IN EXPERIMENTS

Many experiments allow observing the same units repeatedly. We may observe responses from the same person many times and also observe responses to the same items many times. In this section, we examine how this affects our estimation of contrasts between experimental conditions, such as differences in means between treatment and control, i.e., the *average treatment effect* (ATE).<sup>1</sup>

Recent work in applied econometrics has been concerned with dependence due to clustering in data. It is now routine for work in empirical economics to consider and account for dependence in observations produced by one or more types of units. This is reflected in the fact that a recent paper by Cameron et al. [6] on dealing with dependence due to observing two or more types of units repeatedly has been cited over 550 times.<sup>2</sup> Concerns about such dependence have been

<sup>1</sup>There are likely other sources of dependence among observations in online experiments, including some arising from general equilibria in advertising auctions, peer effects, and other “spillovers”. In this paper, we restrict our attention to dependence due to repeated observations of the same units, for which we have inferential procedures, while these other sources of dependence take us into active areas of research beyond the scope of this paper.

<sup>2</sup>Citation count according to Google Scholar [2013-02-22].

featured centrally in methodological work in the context of a growing number of field experiments in economics and other social sciences [12]. Similarly, work on two-way and tensor data in the context of recommender systems and observational comparisons has emphasized the importance of accounting for multiway dependency [18, 19]. And in psychometrics [5] and psycholinguistics [2], investigators have identified problems with ignoring either of two sources of dependence.

As practitioners conducting and analyzing massive Internet experiments, the degree of attention given to this area suggests a need to consider the consequences of dependence for our data. We present our effort to understand whether it would be necessary, in order to have inferential procedures with good performance, to account for *multiple* units causing dependence in our data, or whether a single unit would suffice.

### 2.1 Random effects model

We use the random effects model to illustrate how dependence can affect uncertainty in ATEs, and motivate the use of the bootstrap. Random effects models provide a natural and general way to describe outcomes for data generated by combinations of units, in which each unit and each combination of units contributes a random effect. In the two-way crossed random effects model [2, 24], each observation is generated by some function  $f$  of a linear combination of a grand mean,  $\mu$ , a random effect  $\alpha_i$  for the first variable, which (without loss of generality) we take to be the idiosyncratic deviation for user  $i$ , and a second random variable  $\beta_j$  for the idiosyncratic deviation for item  $j$  (e.g. an ad, a search result, a URL). Finally, we have an error term  $\varepsilon_{ij}$  for each user’s idiosyncratic response to each item.<sup>3</sup> This final term could be caused by a number of factors, including how relevant the item is to the user. Thus, we have the model

$$Y_{ij} = f(\mu + \alpha_i + \beta_j + \varepsilon_{ij})$$

$$\alpha_i \sim \mathcal{H}(0, \sigma_{\alpha_i}^2), \quad \beta_j \sim \mathcal{H}(0, \sigma_{\beta_j}^2), \quad \varepsilon_{ij} \sim \mathcal{H}(0, \sigma_{\varepsilon_{ij}}^2).$$

Each random effect is modeled as being drawn from some distribution  $\mathcal{H}$  with zero mean and some variance. In the homogeneous random effects model, this variance is the same for each user or item (i.e.,  $\sigma_{\alpha_i} = \sigma_{\alpha}$ ), whereas in a heterogeneous random effects model, each variable or groups of variables as may have their own variances.

#### 2.1.1 Comparisons of means

We extend the basic random effects above to consider multiple experimental conditions and develop expressions for the variance of a difference in means between experimental conditions. This illustrates how repeatedly observing the same units, and which units are randomly assigned to conditions, affects this variance.

Let, without loss of generality, users (rather than items) be assigned to experimental conditions. That is, all observations of the same user have the same experimental conditions, such that  $D_{ij} = D_{ij'}$  for all  $j \neq j'$ . As in other work on random effects models where we observe only a small number of combinations of units [18, 19], we work conditional on  $D$ .

<sup>3</sup>For expository simplicity, we consider only a single observation of each user-item pair. An addition error term can be included when there are repeated observations of pairs.

For the sake of exposition, we restrict our attention to linear models with normally distributed random effects. That is, the following analysis considers cases where  $Y$  is unbounded,  $f$  is the identity function, and random effects are drawn from a multivariate normal distribution, so that

$$Y_{ij}^{(d)} = \mu^{(d)} + \alpha_i^{(d)} + \beta_j^{(d)} + \varepsilon_{ij}^{(d)} \\ \vec{\alpha}_i \sim \mathcal{N}(0, \Sigma_\alpha), \quad \vec{\beta}_j \sim \mathcal{N}(0, \Sigma_\beta), \quad \vec{\varepsilon}_{ij} \sim \mathcal{N}(0, \Sigma_\varepsilon). \quad (1)$$

Note that  $\vec{\alpha}_i$ , etc., are vectors, where each element corresponds to the random effect of a unit under a given treatment.

We wish to estimate quantities comparing outcomes for different values of  $D_{ij}$  — most simply, the difference in means, or *average treatment effect* (ATE) for a binary treatment

$$\delta \equiv \mathbb{E}[Y_{ij}^{(1)} \mid D_{ij} = 1] - \mathbb{E}[Y_{ij}^{(0)} \mid D_{ij} = 0] \\ = \mu^{(1)} - \mu^{(0)}.$$

While this difference cannot be directly observed from the data (since user–item pairs can only be assigned to one condition at a time), we can estimate  $\delta$  with the difference in sample means<sup>4</sup> [21]. Our focus is then to consider the true variance of this estimator of  $\delta$  and, later, bootstrap methods for estimating that variance.

The sample mean for each condition is

$$\bar{Y}^{(d)} = \mu^{(d)} + \frac{1}{n_{\bullet\bullet}^{(d)}} \left( \sum_i n_{i\bullet}^{(d)} \alpha_i^{(d)} + \sum_j n_{\bullet j}^{(d)} \beta_j^{(d)} + \sum_i \sum_j \varepsilon_{ij}^{(d)} \right)$$

where, e.g.,  $n_{i\bullet}^{(d)}$  is the number of observations of user  $i$  in condition  $d$ . We then estimate the ATE with  $\hat{\delta} = \bar{Y}^{(1)} - \bar{Y}^{(0)}$ .

Consider the case where the treatment and control groups are of equal size such that  $N = n_{\bullet\bullet}^{(1)} = n_{\bullet\bullet}^{(0)}$ . This enables simplifying the expression for  $\hat{\delta}$  and its variance to

$$\hat{\delta} = \delta + \frac{1}{N} \left[ \sum_i \left( (n_{i\bullet}^{(1)})^2 \alpha_i^{(1)} - (n_{i\bullet}^{(0)})^2 \alpha_i^{(0)} \right) + \sum_j \left( (n_{\bullet j}^{(1)})^2 \beta_j^{(1)} - (n_{\bullet j}^{(0)})^2 \beta_j^{(0)} \right) + \sum_i \sum_j \varepsilon_{ij}^{(D_{ij})} \right]$$

and

$$V[\hat{\delta}] = \frac{1}{N^2} \left[ \sum_i \left( (n_{i\bullet}^{(1)})^2 \sigma_{\alpha^{(1)}}^2 + (n_{i\bullet}^{(0)})^2 \sigma_{\alpha^{(0)}}^2 \right) + \sum_j \left( (n_{\bullet j}^{(1)})^2 \sigma_{\beta^{(1)}}^2 + (n_{\bullet j}^{(0)})^2 \sigma_{\beta^{(0)}}^2 - 2n_{\bullet j}^{(0)} n_{\bullet j}^{(1)} \sigma_{\beta^{(0)}, \beta^{(1)}}^2 \right) + \sum_i \sum_j \left( (n_{ij}^{(1)})^2 \varepsilon_{ij}^{(1)} - (n_{ij}^{(0)})^2 \varepsilon_{ij}^{(0)} \right) \right]. \quad (2)$$

The first term is the contribution of random effects of users to the variance, and the second is the contribution of the random effects of items. The covariance term, present for items, is absent for users and user–item pairs since each is only observed in either the treatment or control.

<sup>4</sup>For true experiments,  $D$  is randomly assigned, but under some circumstances (i.e. conditional ignorability) treatment effects may be estimated without randomization [7, 17, 21].

To further simplify, we can introduce coefficients measuring how much units are duplicated in the data. Following previous work [18, 19], we define

$$\nu_A^{(d)} \equiv \frac{1}{N} \sum_i (n_{i\bullet}^{(d)})^2 \quad \nu_B^{(d)} \equiv \frac{1}{N} \sum_j (n_{\bullet j}^{(d)})^2,$$

which are the average number of observations sharing the same user (the  $\nu_{AS}$ ) or item (the  $\nu_{BS}$ ) as an observation (including itself). For the units assigned to conditions (in this case, users), either  $n_{i\bullet}^{(0)}$  or  $n_{i\bullet}^{(1)}$  is zero for each  $i$ ; for the non-assigned units (items), we need a measure of this between-condition duplication

$$\omega_B \equiv \frac{1}{N} \sum_j n_{\bullet j}^{(0)} n_{\bullet j}^{(1)}.$$

Under the homogeneous random effects model (1), we can then simplify (2) to

$$V[\hat{\delta}] = \frac{1}{N} \left[ \left( \nu_A^{(1)} \sigma_{\alpha^{(1)}}^2 + \nu_A^{(0)} \sigma_{\alpha^{(0)}}^2 \right) + \left( \nu_B^{(1)} \sigma_{\beta^{(1)}}^2 + \nu_B^{(0)} \sigma_{\beta^{(0)}}^2 - 2\omega_B \sigma_{\beta^{(0)}, \beta^{(1)}}^2 \right) + \sigma_{\varepsilon^{(0)}}^2 + \sigma_{\varepsilon^{(1)}}^2 \right]. \quad (3)$$

This expression makes clear that if the random effects for items in the treatment and control are correlated (as we would usually expect), then an increase in the balance of how often items appear in each condition reduces the variance of the estimated treatment effect.

### 2.1.2 Sharp and non-sharp null hypotheses

Under the *sharp null hypothesis*, the treatment has no average or interaction effects; that is, the outcome for a particular user or item is the same regardless of treatment assignment. In the context of our model, this would mean that the variances of the random effects are equal in both conditions and are perfectly correlated across conditions, such that, in addition to  $\delta = 0$ ,

$$\sigma_{\alpha^{(1)}}^2 = \sigma_{\alpha^{(0)}}^2 = \sigma_{\alpha^{(0)}, \alpha^{(1)}}^2 \\ \sigma_{\beta^{(1)}}^2 = \sigma_{\beta^{(0)}}^2 = \sigma_{\beta^{(0)}, \beta^{(1)}}^2 \\ \sigma_{\varepsilon^{(1)}}^2 = \sigma_{\varepsilon^{(0)}}^2 = \sigma_{\varepsilon^{(0)}, \varepsilon^{(1)}}^2. \quad (4)$$

In this case, only random effects for items that are not balanced across conditions contribute to the variance of our ATE estimate: the contribution a single item  $j$  makes to the variance simplifies to  $(n_{\bullet j}^{(0)} - n_{\bullet j}^{(1)})^2 \sigma_{\beta}^2$ ; that is, it depends only on the squared difference in duplication between treatment and control. It is easy to show that

$$V[\hat{\delta}] = \frac{1}{N} \left[ (\nu_A^{(1)} + \nu_A^{(0)}) \sigma_{\alpha}^2 + \kappa_B \sigma_{\beta}^2 + 2\sigma_{\varepsilon}^2 \right], \quad (5)$$

where  $\kappa_B \equiv \frac{1}{N} \sum_j (n_{\bullet j}^{(0)} - n_{\bullet j}^{(1)})^2$  measures the average between-condition duplication of observations of items. If items, like users, also only appear in either treatment or control, then  $\kappa_B = \nu_B^{(1)} + \nu_B^{(0)}$ , highlighting the resulting symmetry between users' and items' contributions to our uncertainty.

When (4) does not hold, we say that there are interaction effects of the treatment and units; for example, there may be an item–treatment interaction effect.

In addition to deviations from the sharp null due to a non-zero average effect, many experiments in domains like search, ads, and recommender systems can result in imbalance and item–treatment interaction effects. For example, a new recommendation model may show different items to users and present items in more (and less) prominent positions. Compared with a null treatment, these changes would produce a smaller  $\omega_B$  and deviations from (4), including a lack of perfect covariance of treatment and control random effects. We can conceive of other treatments that do not change which items are observed, but make some items more likely to produce a response; this would correspond to deviations from (4) only.

Together these considerations highlight that we need to evaluate tests and confidence intervals under conditions other than the sharp null hypothesis, since the variance of our estimated difference can be substantially larger under other more realistic circumstances in which imbalance and interaction effects exist.

### 2.1.3 Choice of experimental unit

We have so far taken it as given that users are the units assigned to conditions, but the under the random effects model, it is clear that other choices, when possible, can increase precision. More generally, our variance expressions highlight that which units are assigned to conditions determines which units can be expected to contribute most of the uncertainty to our estimates of treatment–control comparisons. This creates an asymmetry in two-way data that is not present in prior work on such dependence [6, 18, 19].

It is common in design of industrial, agricultural, and psychological experiments to carefully consider such assignment schemes, including using between-subjects, within-subjects, mixed designs, and blocking to reduce variance [4] while meeting constraints caused by, e.g., carryover effects.

## 2.2 Bootstrapping dependent data

The bootstrap [10] offers a very general method for characterizing the sampling distribution of a statistic (e.g. a difference in means), and can be used to produce confidence intervals for experimental comparisons for many different data generating processes. The bootstrap distribution of a sample statistic is the distribution of that statistic under resampling [10] or reweighting [22] of the sample. In this section, we describe how the bootstrap can be applied to dependent data. We focus on a version of the bootstrap that uses independent weights, rather than the resampling bootstrap, since it is suitable for use in online (i.e., streaming) computational settings [19, 20].

### 2.2.1 The iid bootstrap

In order to get a confidence interval for some statistic  $t$ , we produce  $R$  replicates of the the statistic,  $t_r^*$ , computed on randomly reweighted versions of the sample. That is, for some replicate  $r \in [1, R]$ , each observation  $Y_{ij}$  is randomly reweighted with weights  $W_{ij}$ . These reweighted samples allow us to estimate features of the sampling distribution of our statistic. We generally have  $W_{ij} \sim \mathcal{G}$  where  $\mathcal{G}$  is some distribution with mean and variance 1, such as Poisson(1) and Uniform{0, 2} [19, §3.3]. Note that in this bootstrap, each individual observation is reweighted independently of other observations, including other observations of the same units. Applied to two-way data, the iid bootstrap can be ex-

pected to underestimate the variance of statistics and thus produce confidence intervals with poor coverage [15].

### 2.2.2 Single-way bootstrap

In the single-way bootstrap, or “block” or “cluster” bootstrap, the analyst chooses a single relevant type of unit (e.g., users) and all observations from the same unit are given the same random weight when reweighting. In other words, taken  $i$  as indexing the chosen type of unit, we have  $W_{ij} = W_{ij'} = u_i$  and  $u_i \sim \mathcal{G}$  for all  $j, j'$ . When the data only has one-way dependency, this procedure produces a bootstrap distribution that gives consistent confidence intervals. When the data has additional dependency structure, it can be anticonservative; we use real data and simulations to examine how poorly it works in practice.

### 2.2.3 Multiway bootstrap

When there are two or more relevant units, analysts can use a bootstrap that reweights all relevant units. Under a more general random effects model than the one presented above, the multiway bootstrap produces variance estimates, and thus confidence intervals, that are mildly conservative [18, 19].<sup>5</sup> The two-way bootstrap has been used for analyzing large online advertising experiments [3].

With two-way data, we have  $W_{ij} = u_i v_j$ , where  $u_i \sim \mathcal{G}$  and  $v_j \sim \mathcal{G}$ . That is, the random weights for an observation is the product of two independently sampled weights assigned to unit  $i$  and unit  $j$ . For example, if in one replicate, user  $i$  gets weight 2 and item  $j$  gets weight 3 then all observations of the pair  $(i, j)$  get weight  $2 \times 3 = 6$  in that replicate. Note that if either unit has a weight of 0, any combination of that unit with another unit will be given weight of 0. This procedure can be generalized to cover  $d$ -way data in a straightforward fashion [19].

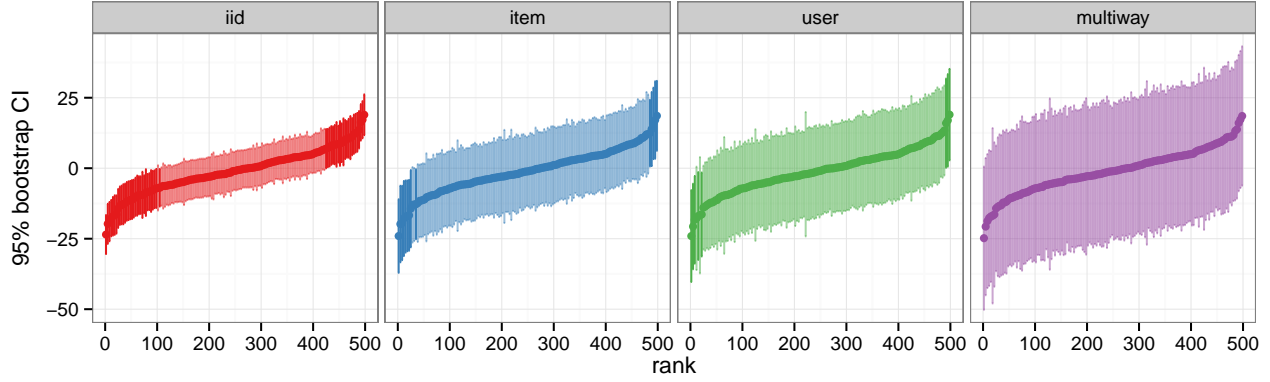
### 2.2.4 Online bootstrapping

For any statistic  $t$  that can be computed online, the single-way bootstrap can be implemented online as follows [19, 20]. On visiting each observation, use a hash of an identifier of each unit (e.g., a user ID) as the seed to the random number generator for  $\mathcal{G}$ , draw  $R$  weights (one for each of the bootstrap replicates), and use these weights to update the running sufficient statistics for  $t_r^*$ . The multiway bootstrap can be implemented online by using the same procedure as for the single-way bootstrap, but at each observation drawing  $R$  weights for each of its  $d$  units and computing their products.

## 2.3 Alternative methods

Bootstrap methods are attractive because they involve minimal assumptions and scale well to large datasets. There are other methods commonly used in practice for statistical inference with dependent data. One could fit a random effects model to the data and then use likelihood-ratio tests or Bayesian inference [2] for the treatment effect parameters of interest. Random effects models require that the experimenter is able to specify a generative model in advance, and require the analyst to make certain assumptions (e.g. homogeneous variances and normality), that are not needed for the bootstrap. Fitting very large crossed random

<sup>5</sup>Previous work in education research [5] and statistics [15] examined the two-way bootstrap for balanced data.



**Figure 1: An illustration of our method for computing true coverage rates for the bootstrap methods with the Search dataset.** We compute 500 A/A tests to obtain nominal “95% confidence intervals” for the difference in means  $\hat{\delta}_{k,r}$ , and count the fraction of tests that accept the null hypothesis (e.g. indicate there is no significant difference in means). To show how results can vary between comparisons, we sort the results by  $E_r[\hat{\delta}_{k,r}]$ , and highlight results that (incorrectly) reject the null. Anti-conservative tests – in this case, the iid and item-clustered bootstrap – reject in more than 5% of the experiments. Differences in the figure are shown relative to the grand mean.

effects models also presents computational difficulties, especially with datasets that span many nodes in a distributed environment.

Recently there has been widespread adoption of cluster robust Huber–White “sandwich” standard errors within econometrics, including extensions to two-way and multiway dependence [6]. These methods are asymptotically consistent for a large class of  $M$ -estimators, but results for these methods are not available for many statistics of interest in large online experiments, such as trimmed and Winsorized means. As with fitting random effects models, sandwich standard errors pose computational difficulties; they require multiple passes through the data and collecting all observations that share a unit, which is not necessary for the bootstrap.

### 3. EMPIRICAL EVALUATION

We evaluate each bootstrap method under random permutations and modifications of the datasets that correspond to various versions of a null hypothesis of no average effect of an experimental treatment on the primary outcome of interest. First, under the *sharp null hypothesis*, the treatment has no effects at all, both on the outcome and on which combination of units are observed. Given our three real datasets, we can produce data consistent with the sharp null simply by randomly assigning units to treatment conditions. Many authors [8, 13, 14] stress the importance of conducting such “A/A tests” as a validation of the combination of one’s random assignment and statistical inference procedures, though it is generally not stated exactly how these null experiments should be carried out or what their limitations are.

#### 3.1 Data

We examine click-through rate outcomes for three core product areas: ads, search, and News Feed. Due to the sensitivity of the data, we only focus on one category of items for each dataset when reporting results from our computational experiments. For example, while there are many different types of items that show up in search results, such

as friends, apps, groups, pages, Web results, etc., the results we present only apply to one of these item types.

**Ads.** We analyze ad click-through rates for one type of ad unit for a popular advertising product on Facebook. Each impression corresponds to a single delivery of the ad to a user’s Web browser.

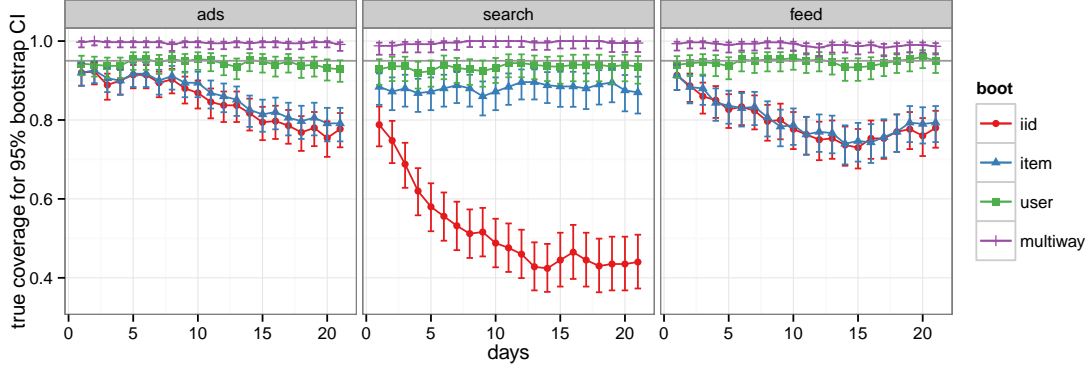
**Search.** We analyze search click-through rates for one type of search result on Facebook. Each impression is a validated delivery of an item in the “typeahead” results, and each click is a click on the item. Note that if an item presented multiple times over several query reformulations, each is considered a separate impression.

**Feed.** We analyze click-through rates for one type of story in the News Feed in a large country. Each impression corresponds to a single delivery of the story to a viewer’s Web browser, and a click corresponds to a click on the item’s thumbnail or snippet.

#### 3.2 Computation

To compute the A/A tests, we first partition the data into  $M$  segments based on the unit we wish to randomize over (i.e. the user ID) such that each segment contains all observations with that corresponding identifier. We then segment the data by taking the identifier of the unit we wish to randomize, concatenating it with a *salt* (i.e., an integer), computing this string’s MD5 hash value,<sup>6</sup> and assigning the unit to a segment number that is integer representation of the first 7 digits of the hashed value modulo  $M$ . In our experiments, we compute the bootstrapped difference in means between every even numbered segment  $m$  and  $m + 1$ , yielding 50 comparisons per salt, and repeat this procedure for 10 salts, yielding  $K = 500$  null experiment comparisons for each method (Figure 1).

<sup>6</sup>Although MD5 is not cryptographically safe (e.g. similar inputs may have correlated outputs), in practice we find that MD5 yields similar results with greater computational efficiency compared to cryptographically safe hashing functions like SHA-1.



**Figure 2: True coverage for nominal 95% confidence intervals produced by the iid, single-way, and multiway bootstrap for A/A tests segmented by user id as a function of time. Uncertainty estimates for the iid and item-level bootstrap become increasingly inaccurate over time, while the user-level and multiway bootstrap have the advertised or conservative Type I error rate.**

	Ads	Search	Feed
users	4,515,816	908,339	545,218
items	317,159	1,362,061	326,831
user-item pairs	24,081,939	4,263,769	2,882,452
$\nu_{\text{users}}$	18.5	35.5	20.3
$\nu_{\text{items}}$	6,625.9	543.6	1,333.0

**Table 1: The amount of duplication present in our datasets for a single 1% segment of users.**

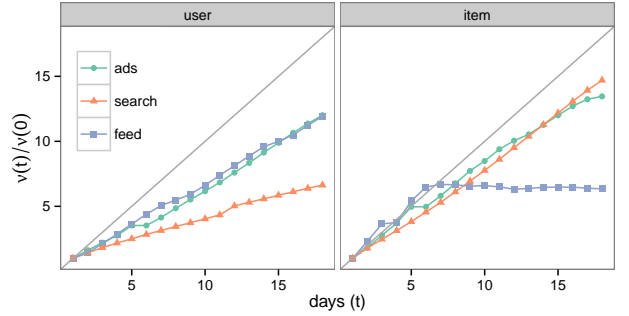
The confidence intervals for each method for each null experiment result from  $R = 500$  bootstrap reweightings of the data. We augment the identifiers of the data using the corresponding 10 salts from the null experiments and apply the bootstrap procedures described in Section 2.2. To determine whether or not a bootstrap experiment is significant, we compute the mean and variance of the difference in means  $\hat{\delta}_{kmr}$  over all  $R$  replicates. The distributions of  $\hat{\delta}_{kmr}$  are asymptotically normal under the bootstrap, so we simply use quantiles of the normal to compute the central  $100(1 - \alpha)\%$  interval.

To obtain the estimated *true coverage* under the sharp null hypothesis (zero mean difference, equal variance), we compute the proportion of times the  $K$  bootstrap tests indicate a significant difference in means at some level  $\alpha$ . We treat each of the  $K$  comparisons as independent, and use the Wilson score interval for binomial proportions [1] to estimate the uncertainty around the coverage.

We may also obtain the coverage for a non-sharp null hypothesis by creating synthetic imbalance between the items in both conditions. To do this, for each pair of segments  $(m, m + 1)$ , we downsample each item from either segment  $m$  or  $m + 1$  (chosen with equal probability); in the downsampled segment for some item  $j$ , its user-item pairs are (independently) removed with probability  $p$ . Thus, when  $p = 0$ , we have the sharp null hypothesis, and when  $p = 1$ , we have total imbalance (i.e., the two conditions contain disjoint sets of items).

### 3.3 Duplication

A central quantity that contributes to the variance of  $\hat{\delta}$



**Figure 3: Duplication ( $\nu$ ) for users and items over time relative to the first day.**

is the average number of observations that share the same user,  $\nu_{\text{user}}$ , and item,  $\nu_{\text{item}}$ . We give basic summary statistics about the duplication in the data for a random 1% segment used in our evaluation for the Ads, Search, and Feed datasets in Table 1. For the restricted categories of items we consider in each dataset, there are more users exposed to ads than the search results or feed stories. While per-user duplication is similar across the three datasets, the per-item duplication for Ads is much higher than either Search or Feed. This pattern is congruent with the nature of the items: the number of businesses that are actively advertising are far fewer than the number of users, while search and News Feed stories tend to have a much longer tail of items that result in lower duplication.

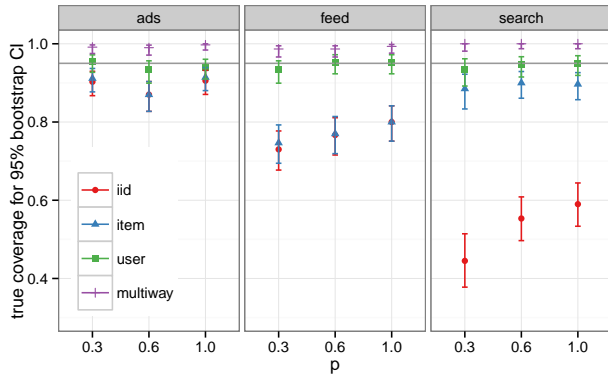
Experiments often run for many days; as the number of days increase, so does the duplication. Figure 3 shows how duplication increases over time. With the exception of Feed items, this relationship is rather linear both for user and items. The behavior for Feed may be explained by the way social media feeds work: unlike ads and search results, users see and interact with very recent content, therefore limiting the average number of users that may be exposed to an item.

Given the two-way random effects model and the increasing relationship between the duplication coefficients and time, we expect that users and items may contribute sub-

stantially to the variance of  $\hat{\delta}$ . Not taking these units into account when computing confidence intervals may result in poor coverage. Figure 2 shows the true coverage of the different bootstrap methods for consecutively larger spans of time in each dataset. We find that the iid confidence intervals tend to be highly anti-conservative. For example, after two weeks of data collection, a search experiment that tests the difference in click-through rates between two equivalent groups of users could result in rejecting the null hypothesis nearly 50% of the time. We find that bootstrapping by the unit not being randomized over (the item) often leads to anti-conservative intervals, and that for the sharp null with little imbalance in items, the user-level bootstrap yields accurate coverage. The multiway bootstrap on the other hand remains conservative no matter how many days are considered.

### 3.4 Imbalance in items

Given how these A/A tests were constructed, there is approximate balance of items across conditions, such that the primary contributors to the variance of  $\hat{\delta}$  are the user and residual error components. However, if items are systematically imbalanced across treatments (e.g. the experiment results in showing similarly relevant, but different ads), then item random effects can also make a substantial contribution (Equation 5). To examine how such imbalance might affect the coverage of the confidence intervals in practice when the treatment has no average or interaction effects, we created imbalance by downsampling items from either condition with probability  $p$  (see Section 3.2 for details).



**Figure 4: True coverage for nominal 95% confidence intervals for each bootstrap method applied to data with varying levels of synthetic imbalance of items across conditions for 2 weeks of data. Imbalance does not appear to affect the accuracy of the true coverage for the multiway and user-level bootstrap, while the iid and item level bootstrap become more conservative when imbalance is greatest.**

Figure 4 shows the true coverage with varying censoring probabilities,  $p \in \{0.3, 0.6, 1.0\}$ . Despite the threat that the imbalance might result in a large item-level contribution to the variance, the coverage of the user bootstrap, which neglects this variance, remains approximately as advertised. This result may be due to a number of factors. First, the most straightforward expressions for  $V[\hat{\delta}]$  and the expected variance estimates from the bootstrap procedures involve as-

suming a homogeneous random effects model, when it can actually be expected that the variances of the random effects for, e.g., frequently observed users are different than those for infrequently observed users. Second, there is a relatively high amount of index-level duplication in the data, such that there are for many users and item a small number of user-item pairs observed; such duplication can cause the multiway bootstrap to be very conservative [19, Theorem 7].

For Feed and Search, the poor coverage of the iid and item bootstrap confidence intervals notably increases, though they continue to undercover. This is expected since, in addition to creating imbalance, the downsampling procedure reduces within-condition duplication.

## 4. SIMULATIONS

We have seen how different bootstrap methods perform under the sharp null hypothesis and synthetic imbalance with three real-world domains. However, these A/A tests cannot tell us about how bootstrap procedures might perform in situations where treatments do have effects. For example, an ads experiment that manipulates the display of certain advertising units may only affect certain items and not others [3]. To explore these circumstances, we conduct simulations with a probit random effects model parameterized to mirror the kinds of outcomes described in the previous section. We use this generative model to vary the presence of an item-treatment interaction, a plausible source of violations of the sharp null hypothesis.

We modify the model of (1) so that  $Y$  is binary and there is a single intercept common to both treatment and control, reflecting the lack of an ATE:

$$y_{ij}^{(d)} = \mu + \alpha_i^{(d)} + \beta_j^{(d)} + \varepsilon_{ij}^{(d)} \quad (6)$$

$$E[Y_{ij}^{(d)}] = \mathbb{1}\{y_{ij}^{(d)} > 0\} \quad (7)$$

Also reflecting the absence of an ATE, we restrict the random effect variance to be the same in treatment and control. For example, the covariance matrix for the item random effects is

$$\Sigma_\beta = \begin{bmatrix} \sigma_\beta^2 & \rho_\beta \sigma_\beta^2 \\ \rho_\beta \sigma_\beta^2 & \sigma_\beta^2 \end{bmatrix}.$$

To make realistic choices for the variances of the random effects, we fit a probit random effects models to the ads dataset from a large random sample of users in each of several small countries. This produced several estimates of  $\sigma_\alpha$  and  $\sigma_\beta$ . We report on simulation results for  $\sigma_\alpha = 0.3$ , which is close to several of the estimates. Our estimates of  $\sigma_\beta$  often ranged from 0.2 to 0.9, so we present results for  $\sigma_\beta \in \{0.1, 0.3, 0.5, 1.0\}$ . We set  $\mu$  so as to achieve  $E[Y_{ij}]$  close to 0.02.<sup>7</sup>

We constructed the set of observed user-item pairs used in the simulations by assigning each of 3,000 potential users and 200 potential ads to log-normally distributed scores. For each of  $2N$  observations, we selected a particular user and ad with probability proportional to this score. This yielded a “layout” with 2481 unique users, 199 unique ads, and duplication coefficients  $\nu_A \doteq 30.9$  and  $\nu_B \doteq 6077.4$ , which is similar to the Ads dataset.

<sup>7</sup>Since there is no scale to the latent variable  $y_{ij}$ , we achieved this by in fact choosing a fixed  $\mu = -2$  and rescaling the random effect variances to sum to 1.



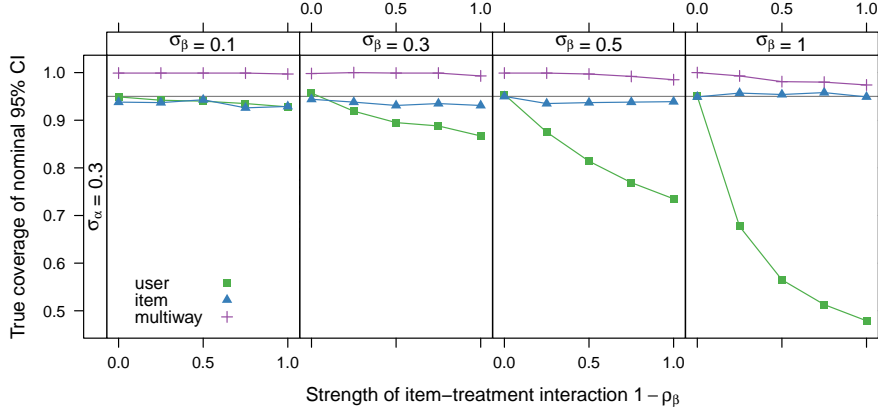


Figure 5: Effects of item-treatment interaction effects on true coverage of 95% confidence intervals. Decreasing  $\rho_\beta$ , which makes the random item effects less correlated between treatment and control, reduces the coverage of user bootstrap confidence intervals. This effect is moderated by the magnitude of the item-level random effects.

#### 4.1 Item-treatment interactions

Even if the treatment has no effects on average, it can have positive effects for some users and items and negative effects for others. Given our random effects model, we know that item-treatment interactions can increase the contribution of duplication of items to the variance of the mean difference.

We vary item-treatment interactions by setting the correlation coefficient  $\rho_\beta \in \{0, 0.25, 0.5, 0.75, 1\}$ . Perfect correlation  $\rho_\beta = 1$  corresponds to the sharp null hypothesis, while decreasing  $\rho_\beta$  corresponding to an increasing proportion of item random effects being not shared across conditions. At the extreme of  $\rho_\beta = 0$ , the random effect of an item in the treatment is completely independent of its random effect in the control.

#### 4.2 Results

Figure 5 summarizes the results of 1000 simulations for each combination of parameter values. Without any item-treatment interaction, both the user and item bootstrap have approximately correct coverage since; this is attributable to the relatively low  $\nu_A$  in this simulation, and is consistent with the results from a small number of days in our real datasets. As the item-treatment interaction increases, the coverage of the user bootstrap confidence intervals drop substantially. For example, even with moderate values of  $\sigma_\beta = 0.5$  and  $\rho_\beta = 0.75$ , a nominally 95% confidence interval has a true coverage of 87.5%. While do we not expect to observe the extremes of all item-level variance being treatment specific (i.e.,  $\rho_\beta = 0$ ), these results demonstrate that deviations from the sharp null in the form of item-treatment interaction have serious consequences for the single-way bootstrap. On the other hand, the multiway bootstrap remains mildly conservative even with large  $\sigma_\beta$  and small  $\rho_\beta$ .

### 5. DISCUSSION

Despite having a large number of individual observations, many settings for online experiments involve substantial de-

pendence and small effects such that statistical inference remains a central concern. The preceding analysis of real and simulated data makes clear that methods which neglect dependence structure in these large experiments can result in high Type I error rates and confidence intervals with poor coverage. In each of our three datasets, the iid bootstrap performed very poorly, such that using it (or other methods assuming iid observations) would result in reaching incorrect conclusions about the presence, sign, and magnitude of treatment effects [11].

On the other hand, neglecting dependence among observations of units not assigned to conditions (the items) generally did not result in lower coverage with our data. For each of the datasets, this remained the case even when we produced imbalance of items across conditions. Given the random effects model posited in Section 2.1, one might expect this imbalance to make both the user and item contributions to the variance necessary to account for separately. Since bootstrapping multiple units and storing these replicates can have substantial costs in terms of computation and infrastructure, our results suggest that experimenters should consider whether a single-way bootstrap on the experimental units may be practically sufficient, even in the presence of other clearly relevant units, such as ads and URLs.

Nonetheless, neglecting dependence among observation of these non-experimental units may have substantial effects on coverage when the treatment has any effects. Most treatments are expected to have some effects. Our simulations with item-treatment interaction effects demonstrate that the coverage of the user bootstrap can be extremely sensitive to the presence of these effects. This highlights that using A/A tests only serves to validate inferential procedures under a narrow set of conditions (i.e., the sharp null hypothesis), but cannot detect other (potentially severe) inferential problems that occur in other circumstances. Given that experimenters expect treatment effects, and often want to know how large the average effects are, they should consider whether or not they wish to use a procedure that provides a somewhat conservative measurement of uncertainty



(i.e. the multiway bootstrap), or the user-level bootstrap, which correctly tests the less plausible sharp null.

A limitation of the present work is that, from the perspective of experimenters such as ourselves trying to evaluate inferential methods in practice, there is remaining gap between what is possible to learn from straightforward perturbations of real datasets and what is possible to learn from necessarily simplified generative models. Future work may develop more sophisticated ways of perturbing existing data and using additional parameters estimated from real experiments to produce evaluations for data that more closely resemble outcomes in the field.

This paper has been primarily concerned with Type I error rates and the coverage of confidence intervals, but experimenters are equally concerned about Type II errors (failures to reject the null) and related errors such as incorrectly estimating the direction or magnitude of effects. Many principled approaches to choosing how to assign units to one of many available treatments over time (e.g. solutions to multi-armed bandit problems) require correctly estimating one's uncertainty about the expected payoffs of the treatments [23]. Therefore, we expect that addressing multiway dependence will remain important when taking these approaches as well. A related point is that experimenters often exert considerable effort *reducing* the width of CIs by increasing precision through design and adjustment [4, 9, 16]. Many of these methods could be applied in combination with single or multiway bootstrapping. Finally, there may other practical ways to reduce the width of multiway bootstrap CIs through using linear combinations of variance estimates from different bootstrap procedures [6, 19].

## 6. ACKNOWLEDGEMENTS

We would like to thank Daniel Ting, Wojciech Galuba, and Art Owen for their thoughtful comments.

## 7. REFERENCES

- [1] A. Agresti. *Categorical Data Analysis*. Wiley-Interscience, 2nd edition, July 2002.
- [2] R. H. Baayen, D. J. Davidson, and D. M. Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412, 2008.
- [3] E. Bakshy, D. Eckles, R. Yan, and I. Rosenn. Social influence in social advertising: evidence from field experiments. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 146–161. ACM, 2012.
- [4] G. E. Box, J. S. Hunter, and W. G. Hunter. *Statistics for experimenters: design, innovation, and discovery*, volume 13. Wiley Online Library, 2005.
- [5] R. L. Brennan, D. J. Harris, and B. A. Hanson. *The bootstrap and other procedures for examining the variability of estimated variance components in testing contexts*. American College Testing Program, 1987.
- [6] A. Cameron, J. Gelbach, and D. Miller. Robust inference with multi-way clustering. *Journal of Business & Economic Statistics*, 29(2):238–249, 2011.
- [7] D. Chan, R. Ge, O. Gershony, T. Hesterberg, and D. Lambert. Evaluating online ad campaigns in a pipeline: causal models at scale. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 7–16. ACM, 2010.
- [8] T. Crook, B. Frasca, R. Kohavi, and R. Longbotham. Seven pitfalls to avoid when running controlled experiments on the web. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1105–1114. ACM, 2009.
- [9] A. Deng, Y. Xu, R. Kohavi, and T. Walker. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 123–132. ACM, 2013.
- [10] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.
- [11] A. Gelman and F. Tuerlinckx. Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, 15(3):373–390, 2000.
- [12] A. S. Gerber and D. P. Green. *Field Experiments: Design, Analysis, and Interpretation*. WW Norton, 2012.
- [13] R. Kohavi, A. Deng, B. Frasca, R. Longbotham, T. Walker, and Y. Xu. Trustworthy online controlled experiments: five puzzling outcomes explained. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 786–794. ACM, 2012.
- [14] R. Kohavi, R. Longbotham, D. Sommerfield, and R. Henne. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery*, 18(1):140–181, 2009.
- [15] P. McCullagh. Resampling and exchangeable arrays. *Bernoulli*, 6(2):285–301, 2000.
- [16] L. W. Miratrix, J. S. Sekhon, and B. Yu. Adjusting treatment effect estimates by post-stratification in randomized experiments. *JR Stat. Soc. Ser. B. Stat. Methodol.* To appear, 2012.
- [17] S. L. Morgan and C. Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, July 2007.
- [18] A. B. Owen. The pigeonhole bootstrap. *The Annals of Applied Statistics*, 1(2):386–411, 2007.
- [19] A. B. Owen and D. Eckles. Bootstrapping data arrays of arbitrary order. *The Annals of Applied Statistics*, 6(3):895–927, 2012.
- [20] N. C. Oza and S. Russell. Experimental comparisons of online and batch versions of bagging and boosting. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 359–364. ACM, 2001.
- [21] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.
- [22] D. B. Rubin. The Bayesian bootstrap. *The Annals of Statistics*, 9(1):130–134, 1981.
- [23] S. L. Scott. A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010.
- [24] S. R. Searle, G. Casella, C. E. McCulloch, et al. *Variance Components*. Wiley New York, 1992.